derivative of cost function for Logistic Regression

Asked 7 years, 10 months ago Active 1 year, 4 months ago Viewed 88k times

112

112

I am going over the lectures on Machine Learning at Coursera.

I am struggling with the following. How can the partial derivative of

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))$$

where $h_\theta(x)$ is defined as follows

$$h_\theta(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1+e^{-z}}$$

be

$$\frac{\partial}{\partial\theta_j} J(\theta) = \sum_{i=1}^{m} (h_\theta(x^i) - y^i)x_j^i$$

In other words, how would we go about calculating the partial derivative with respect to θ of the cost function (the logs are natural logarithms):

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))$$

statistics regression machine-learning partial-derivative

Share Cite Follow

edited Aug 27 '13 at 12:26
Avitus
13k 1 23 45

asked Aug 27 '13 at 10:41
dreamwalker
1,245 3 9 6

I think to resolve θ by gradient will be hard way (or impossible??). Because it different with linear classfication, it will not has close form. So i suggest you can use other method example Newton's method. BTW, do you find θ using above way? – John Jul 22 '14 at 2:16

5 missing $\frac{1}{m}$ for the derivative of the Cost – bourneli Apr 20 '17 at 5:01

7 Answers

Active Oldest Votes

160

The reason is the following. We use the notation:

$$\theta x^i := \theta_0 + \theta_1 x_1^i + \cdots + \theta_p x_p^i.$$

Then

$$\log h_\theta(x^i) = \log \frac{1}{1+e^{-\theta x^i}} = -\log(1+e^{-\theta x^i}),$$

$$\log(1-h_\theta(x^i)) = \log(1-\frac{1}{1+e^{-\theta x^i}}) = \log(e^{-\theta x^i}) - \log(1+e^{-\theta x^i}) = -\theta x^i - \log(1+e^{-\theta x^i}),$$

[this used: $1 = \frac{(1+e^{-\theta x^i})}{(1+e^{-\theta x^i})}$, the 1's in numerator cancel, then we used: $\log(x/y) = \log(x) - \log(y)$]

Since our original cost function is the form of:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^m y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))$$

Plugging in the two simplified expressions above, we obtain

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^m \left[-y^i(\log(1+e^{-\theta x^i})) + (1-y^i)(-\theta x^i - \log(1+e^{-\theta x^i}))\right]$$

, which can be simplified to:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^m \left[y_i\theta x^i - \theta x^i - \log(1+e^{-\theta x^i})\right] = -\frac{1}{m}\sum_{i=1}^m \left[y_i\theta x^i - \log(1+e^{\theta x^i})\right], \quad (*)$$

where the second equality follows from

$$-\theta x^i - \log(1+e^{-\theta x^i}) = -\left[\log e^{\theta x^i} + \log(1+e^{-\theta x^i})\right] = -\log(1+e^{\theta x^i}).$$

[we used $\log(x) + \log(y) = log(xy)$]

All you need now is to compute the partial derivatives of $(*)$ w.r.t. θ_j. As

$$\frac{\partial}{\partial \theta_j} y_i \theta x^i = y_i x^i_j,$$

$$\frac{\partial}{\partial \theta_j} \log(1+e^{\theta x^i}) = \frac{x^i_j e^{\theta x^i}}{1+e^{\theta x^i}} = x^i_j h_\theta(x^i),$$

the thesis follows.

Share Cite Follow

edited Jan 13 '19 at 14:45
amWhy
201k 133 257 475

answered Aug 27 '13 at 12:25
Avitus
13k 1 23 45

1 Can't upvote as I don't have 15 reputation just yet! :) Will google the maximum entropy principle as I have no clue what that is! as a side note I am not sure how you made the jump from log(1 - hypothesis(x)) to log(a) - log(b) but will raise another question for this as I don't think I can type latex here, really impressed with your answer! learning all this stuff on my own is proving to be quite a challenge thus the more kudos to you for providing such an elegant answer! :) – dreamwalker Aug 27 '13 at 13:54

1 yes!!! I couldn't see that you were using this property $\log(\frac{a}{b}) = \log a - \log b$ Now everything makes sense :) Thank you so much! :) – dreamwalker Aug 27 '13 at 14:26

5 Awesome explanation, thank you very much! The only thing I am still struggling with is the very last line, how the derivative was made in

$$\frac{\partial}{\partial \theta_j} \log(1+e^{\theta x^i}) = \frac{x^i_j e^{\theta x^i}}{1+e^{\theta x^i}}$$

? Could you provide a hint for it? Thank you very much for the help! – Pedro Lopes Dec 1 '15 at 21:40

10 @codewarrior hope this helps.

$$\frac{\partial}{\partial \theta_j} \log(1+e^{\theta x^i}) = \frac{x^i_j e^{\theta x^i}}{1+e^{\theta x^i}}$$

$$= \frac{x^i_j}{e^{-\theta x^i} * (1+e^{\theta x^i})}$$

$$= \frac{x^i_j}{e^{-\theta x^i} + e^{-\theta x^i + \theta x^i}}$$

$$= \frac{x^i_j}{e^{-\theta x^i} + e^0}$$

$$= \frac{x^i_j}{e^{-\theta x^i} + 1}$$

$$= \frac{x^i_j}{1 + e^{-\theta x^i}}$$

$$= x_j^i * h_\theta(x^i)$$

as

$$h_\theta(x^i) = \frac{1}{1 + e^{\theta x^i}}$$

– Rudresha Parameshappa Jan 2 '17 at 13:06

2 @Israel, logarithm is usually base e in math. Take a look at [When log is written without a base, is the equation normally referring to log base 10 or natural log?](#) – gdrt Mar 11 '18 at 11:46

4

@pedro-lopes, it is called as: [chain rule](#).

$$(u(v))' = u(v)' * v'$$

For example:

$$y = \sin(3x - 5)$$

$$u(v) = \sin(3x - 5)$$

$$v = (3x - 5)$$

$$y' = \sin(3x - 5)' = \cos(3x - 5) * (3 - 0) = 3\cos(3x - 5)$$

Regarding:

$$\frac{\partial}{\partial \theta_j} \log(1 + e^{\theta x^i}) = \frac{x_j^i e^{\theta x^i}}{1 + e^{\theta x^i}}$$

$$u(v) = \log(1 + e^{\theta x^i})$$

$$v = 1 + e^{\theta x^i}$$

$$\frac{\partial}{\partial \theta} \log(1 + e^{\theta x^i}) = \frac{\partial}{\partial \theta} \log(1 + e^{\theta x^i}) * \frac{\partial}{\partial \theta}(1 + e^{\theta x^i}) = \frac{1}{1 + e^{\theta x^i}} * (0 + x e^{\theta x^i}) = \frac{x e^{\theta x^i}}{1 + e^{\theta x^i}}$$

Note that

$$\log(x)' = \frac{1}{x}$$

Hope that I answered on your question!

Share Cite Follow

edited Jun 12 '20 at 10:38
Community ♦
1

answered Apr 17 '17 at 13:17
RedEyed
141 4

4

We have,

$$L(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y_i.\, logP(y_i|x_i,\theta) + (1-y_i).\log(1-P(y_i|x_i,\theta))$$

$$h_\theta(x_i) = P(y_i|x_i,\theta) = P(y_i=1|x_i,\theta) = \frac{1}{1+\exp\left(-\sum_k \theta_k x_i^k\right)}$$

Then,

$$\log(P(y_i|x_i,\theta)) = \log(P(y_i=1|x_i,\theta)) = -\log\left(1+\exp\left(-\sum_k \theta_k x_i^k\right)\right)$$

$$\Rightarrow \frac{\partial}{\partial\theta_j} logP(y_i|x_i,\theta) = \frac{x_i^j.\exp\left(-\sum_k \theta_k x_i^k\right)}{1+\exp\left(-\sum_k \theta_k x_i^k\right)} = x_i^j.(1-P(y_i|x_i,\theta))$$

and

$$\log(1-P(y_i|x_i,\theta)) = \log(1-P(y_i=1|x_i,\theta)) = -\sum_k \theta_k x_i^k - \log\left(1+\exp\left(-\sum_k \theta_k x_i^k\right)\right)$$

$$\Rightarrow \frac{\partial}{\partial\theta_j}\log(1-P(y_i|x_i,\theta)) = -x_i^j + x_i^j.(1-P(y_i|x_i,\theta)) = -x_i^j.P(y_i|x_i,$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial\theta_j}L(\theta) &= -\frac{1}{m}\sum_{i=1}^{m} y_i.\frac{\partial}{\partial\theta_j}logP(y_i|x_i,\theta) + (1-y_i).\frac{\partial}{\partial\theta_j}\log(1-P(y_i|x_i,\theta)) \\ &= -\frac{1}{m}\sum_{i=1}^{m} y_i.x_i^j.(1-P(y_i|x_i,\theta)) - (1-y_i).x_i^j.P(y_i|x_i,\theta) \\ &= -\frac{1}{m}\sum_{i=1}^{m} y_i.x_i^j - x_i^j.P(y_i|x_i,\theta) \\ &= \frac{1}{m}\sum_{i=1}^{m}(P(y_i|x_i,\theta)-y_i).x_i^j\end{aligned}$$

(Proved)

Share Cite Follow edited Dec 5 '17 at 11:42 answered Nov 27 '17 at 12:50

Sandipan Dey
1,054 6 9

The logistic regression implementation with gradient-descent using this derivative can be found here: sandipanweb.wordpress.com/2017/11/25/... – Sandipan Dey Nov 27 '17 at 12:53

what about w.r.t to b? – user_6396 Jul 15 '19 at 2:59

1 We can include the bias term θ_0 inside θ if we extend x_i as $(1, x_i)$, i.e., by adding a column of 1 s with x. – Sandipan Dey Jul 15 '19 at 7:34

3

Pedro, => partial fractions

$$\log(1 - \frac{a}{b})$$

$$1 - \frac{a}{b} = \frac{b}{b} - \frac{a}{b} = \frac{b-a}{b},$$

$$\log(1 - \frac{a}{b}) = \log(\frac{b-a}{b}) = \log(b-a) - \log(b)$$

Share Cite Follow

edited Apr 13 '16 at 15:39

answered Apr 13 '16 at 15:23

Richard Wheatley

41 ▲3

3

You have to get the partial derivative with respect θ_j. Remember that the hypothesis function here is equal to the sigmoid function which is a function of θ; in other words, we need to apply the chain rule. This is my approach:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))$$

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{\partial}{\partial\theta_j}[-\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))]$$

Anything without θ is treated as constant:

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \frac{\partial}{\partial\theta_j}[\log(h_\theta(x^i))] + (1-y^i)\frac{\partial}{\partial\theta_j}[\log(1-h_\theta(x^i))] \tag{1}$$

Let's solve each derivative separately and then plug back in on (1):

$$\frac{\partial}{\partial\theta_j}[\log(h_\theta(x^i))] = \frac{1}{h_\theta(x^i)}\frac{\partial}{\partial\theta_j}h_\theta(x^i) \tag{2}$$

$$\frac{\partial}{\partial\theta_j}[\log(1-h_\theta(x^i))] = \frac{1}{1-h_\theta(x^i)}\frac{\partial}{\partial\theta_j}(1-h_\theta(x^i) = \frac{-1}{1-h_\theta(x^i)}\frac{\partial}{\partial\theta_j}h_\theta(x^i) \tag{3}$$

Plug (3) and (2) in (1):

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \frac{1}{h_\theta(x^i)}\frac{\partial}{\partial\theta_j}h_\theta(x^i) + (1-y^i)\frac{-1}{1-h_\theta(x^i)}\frac{\partial}{\partial\theta_j}h_\theta(x^i)]$$

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[\frac{y^i}{h_\theta(x^i)} - \frac{(1-y^i)}{1-h_\theta(x^i)}] * \frac{\partial}{\partial\theta_j}h_\theta(x^i) \tag{4}$$

Notice that using the chain rule, the derivative of the hypothesis function can be understood as

$$\frac{\partial}{\partial\theta_j}[\ h_\theta(x^i)] = \frac{\partial}{\partial z}[\ h(z)] * \frac{\partial}{\partial\theta_j}[\ z(\theta)] = [h(z) * [1-h(z)]] * [x_j^i] \tag{5}$$

where

$$\frac{\partial}{\partial z}[\ h(z)] = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = \frac{0-(1)*(1+e^{-z})'}{(1+e^{-z})^2} = \frac{(e^{-z})}{(1+e^{-z})^2} = [\frac{1}{(1+e^{-z})}]$$
$$*[\frac{(e^{-z})}{(1+e^{-z})}] = [\frac{1}{(1+e^{-z})}]*[1-\frac{1}{(1+e^{-z})}] = h(z)*[1-h(z)]$$

and

$$\frac{\partial}{\partial\theta_j}[\ z(\theta)] = \frac{\partial}{\partial\theta_j}[\ \theta x^i] = x^i_j$$

Plug (5) in (4):

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[\frac{y^i}{h_\theta(x^i)} - \frac{(1-y^i)}{1-h_\theta(x^i)}]*[h_\theta(x^i)*(1-h_\theta(x^i))*x^i_j]$$

Applying some algebra and solving subtraction:

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i)-y^i)x^i_j$$

There is a $1/m$ factor missing on your expected answer.

Hope this helps.

Share Cite Follow

edited Feb 9 '20 at 20:37
Stefan Egger
697 2 12

answered Feb 9 '20 at 20:05
Victor Arango
31 1

2

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i\log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))$$

where $h_\theta(x)$ is defined as follows

$$h_\theta(x) = g(\theta^T x),$$

$$g(z) = \frac{1}{1+e^{-z}}$$

Note that $g(z)' = g(z)*(1-g(z))$ and we can simply write right side of summation as

$$y\log(g) + (1-y)\log(1-g)$$

and the derivative of it as

$$
\begin{aligned}
& y\frac{1}{g}g' + (1-y)\left(\frac{1}{1-g}\right)(-g') \\
&= \left(\frac{y}{g} - \frac{1-y}{1-g}\right)g' \\
&= \frac{y(1-g) - g(1-y)}{g(1-g)}g' \\
&= \frac{y - y*g - g + g*y}{g(1-g)}g' \\
&= \frac{y - y*g - g + g*y}{g(1-g)}g(1-g)*x \\
&= (y-g)*x
\end{aligned}
$$

and then we can rewrite above as

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_j^i$$

Share Cite Follow

answered Dec 5 '18 at 11:59
agile
21 ▲1

In your derivation, from where did you get "x". I mean , you replaced g' with g(1-g)*x but g' = g(1-g) right ? from where "x" come into picture – Ravi Kumar B Oct 4 '19 at 4:17

1

Notice that,

$$\frac{\partial}{\partial\theta_j}y_i\theta x^i = \frac{\partial}{\partial\theta_j}y_i(\theta_0 + \theta_1 x_1^i + ... + \theta_j x_j^i) =$$

in this $\partial\theta_j$ order derivative, y_i is a constant, so

$$= y_i\frac{\partial}{\partial\theta_j}(\theta_0 + \theta_1 x_1^i + ... + \theta_j x_j^i) =$$

because it is a linear model ($\frac{\partial}{\partial\theta}k\theta = k$), so

$$= y_i(0 + x_1^i + ... + x_j^i) =$$

$$= y_i x_j^i$$

Finally,

$$\frac{\partial}{\partial\theta_j}y_i\theta x^i = y_i x_j^i$$

Share Cite Follow

answered Mar 24 '19 at 13:12
TSRTSR
11 ▲1